MAP-MUSIC2VEC: A SIMPLE AND EFFECTIVE BASELINE FOR SELF-SUPERVISED MUSIC AUDIO REPRESENTATION LEARNING

Yizhi Li[1]*, Ruibin Yuan[2,4]*, Ge Zhang[2,5]*, Yinghao Ma[3]*, Chenghua Lin[1]†, Xingran Chen[5], Anton Ragni[1], Hanzhi Yin[4], Zhijie Hu[6], Haoyu He[7], Emmanouil Benetos[3], Norbert Gyenge[1], Ruibo Liu[8], Jie Fu[2]†

[1]Department of Computer Science, University of Sheffield, UK { yizhi.li, c.lin } @sheffield.ac.uk; [2]Beijing Academy of Artificial Intelligence, China fujie@baai.ac.cn; [3]Centre for Digital Music, Queen Mary University of London, UK; [4]School of Music, Carnegie Mellon University, PA, USA; [5]University of Michigan Ann Arbor, USA; [6]HSBC Business School, Peking University, China; [7]University of T¨ubingen & MPI-IS, Germany; [8]Dartmouth College, NH, USA.

1. Introduction

Existing music self-supervised learning models (e.g., Jukebox [1]) are expensive to finetune, though the results are impressive on music information retrieval tasks [2].

- Designing **MAP-Music2Vec** following the principles proposed in data2vec [3].
- Less than 2% of the parameters of the Jukebox, and therefore, trainable in a single GPU.
- Achieving comparable results to Jukebox.
- The model will be released on Huggingface.

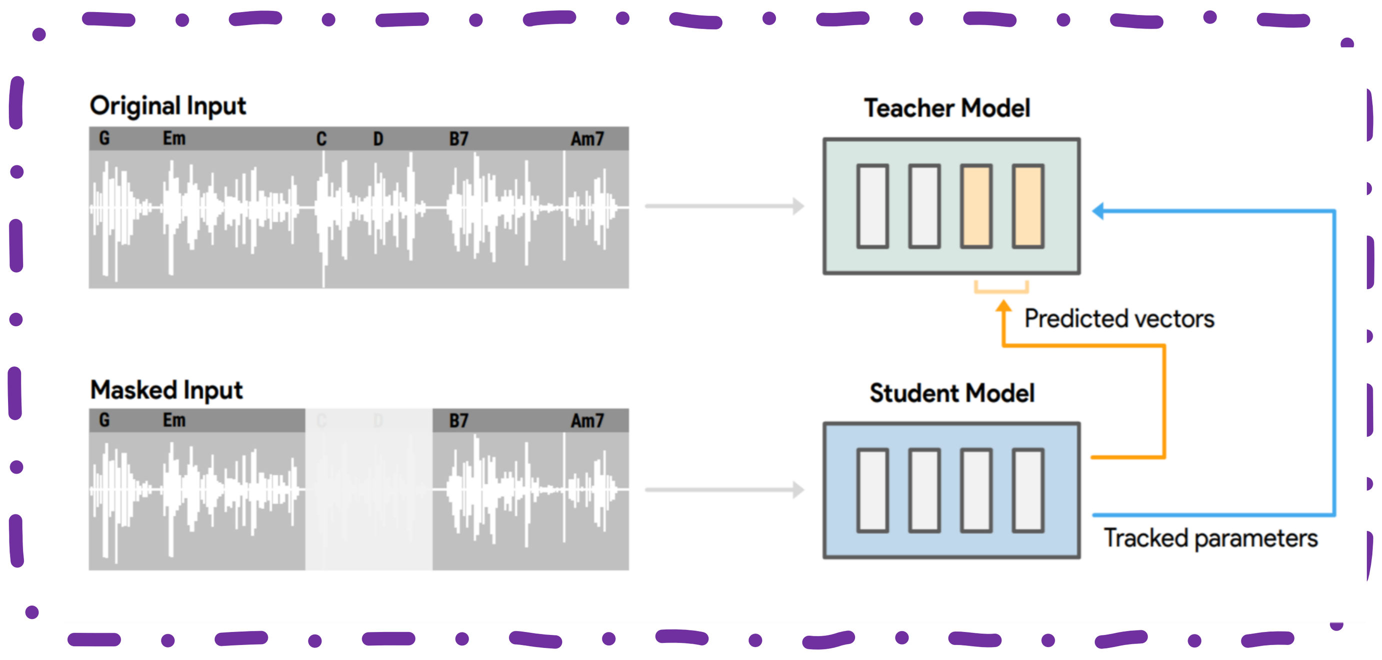

Figure 1: Music2Vec Framework.
During pre-training: the student model aims to reconstruct the masked music audios by taking the contextualised representations provided by the teacher model as a prediction target.

2. Methodology

- A teacher model in the same architecture is used to provide prediction targets.
- The teacher is updated according to the exponential moving average of the student.
- The student takes the masked input and predicts the average of top-K layer outputs of the teacher model, which takes the unmasked input.
- The encodes uses a multi-layer 1-D CNN feature extractor, and further input these tokens to a 12-layer Transformer.
- We trained Music2Vec models on 1k hours of 30s music audio, with 8 × NVIDIA A100-40GB GPUs around 6 days for 400k steps.

3. Pre-training Experiments

- **Audio Length Cropping**
- **Mask Strategy**
- **Prediction Target Layer**

	Approach		Tags (MTT [6]) AUC	Tags (MTT [6]) AP	Genre (GTZAN [7]) Accuracy	Key(GS [8]) Score	Emotion (EMO [9]) $R2_{Arousal}$	Emotion (EMO [9]) $R2_{Valence}$	Average
Baselines	CHOI [10]		89.7	36.4	75.9	13.1	67.3	43.4	51.9
Baselines	MUSICNN [11]		90.6	38.3	79.0	12.8	70.3	46.6	53.7
Baselines	CLMR [12]		89.4	36.1	68.6	14.9	67.8	45.8	50.8
Baselines	Jukebox [1]		<u>91.5</u>	<u>41.4</u>	<u>79.7</u>	<u>66.7</u>	<u>72.1</u>	<u>61.7</u>	<u>69.9</u>
Music2Vec	Starting Setting		88.2	34.1	61.7	32.1◇	66.2	45.8	54.7
Music2Vec	Length Crop	5s	$\boxed{89.5}$	35.9	$\boxed{76.6}$	50.1◇	$\boxed{69.4}$	$\boxed{57.4}$	$\boxed{63.2}$
Music2Vec	Length Crop	10s	89.0	$\boxed{36.0}$	70.3	27.4	62.7	46.1	55.3
Music2Vec	Length Crop	15s	88.3	34.1	65.9	38.1◇	60.1	43.6	55.0
Music2Vec	Mask Span	5	87.0◇	32.2◇	59.3	29.5	50.3◇	24.7◇	47.2
Music2Vec	Mask Span	15	87.8	33.3	65.2	41.9◇	55.0	36.9	53.4
Music2Vec	Mask Prob	50%	87.7	33.2	62.8	43.6◇	54.8	37.6	53.2
Music2Vec	Mask Prob	70%	87.2	32.4	60.7	35.3◇	55.3◇	36.0◇	51.2
Music2Vec	Mask Prob	80%	87.5	32.7	60.0◇	34.6◇	50.7	40.4	51.0
Music2Vec	Target Step	Top-12	88.8	34.5	65.2	$\boxed{50.8}$◇	67.4	43.8	58.4
Music2Vec	Target Step	800K	87.6◇	33.2◇	60.3	44.9◇	54.8◇	40.8◇	53.6

These are *probing* results. We could further *fine-tune* Music2Vec to achieve better performance.

Table 2: Overall Results of Self-Supervised Models.
Underline and square boxes indicate the best overall performance and the best setting of Music2Vec, respectively. ◇ indicates the results are produced by the convolutional feature extractor representations. Results of baselines are taken from JukeMIR.

Model Release

References

[1] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” arXiv preprint arXiv:2005.00341, 2020.
[2] R. Castellon, C. Donahue, and P. Liang, “Codified audio language modeling learns useful representations for music information retrieval,” in Proc. of International Conference on Music Information Retrieval (ISMIR), 2021.
[3] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “Data2vec: A general framework for selfsupervised learning in speech, vision and language,” arXiv preprint arXiv:2202.03555, 2022.

Acknowledgements

This paper is a tribute to our talented friend Anqiao Yang, for his friendship and valuable advice to this work.
Yizhi Li is fully funded by an industrial PhD studentship (Grant number: 171362) from the University of Sheffield, UK.
Yinghao Ma is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported by UK Research and Innovation [grant number EP/S022694/1].
This work is supported by the National Key R&D Program of China (2020AAA0105200).
We acknowledge IT Services at The University of Sheffield for the provision of services for High Performance Computing.
We would also like to express great appreciation for the suggestions from faculties Dr Chris Donahue, and Dr Roger Dannenberg, as well as the facility support from Mr. Yulong Zhang in the preliminary stage.

4. Results

- Following the probing settings in JukeMIR [2], we evaluate tasks including music tagging, genre classification, key detection, and emotion recognition.
- MAP-Music2Vec achieves comparable results to Jukebox with less than 1/50 parameters.
- The music recording length is negatively correlated to the Music2Vec performances, which suggests that our model relies too much on local information.
- The representations of CNN extractor sometimes outperform the Transformer layers, especially for key detection.
- Increasing the training steps, changing the mask span, or changing the mask probability does not give performance gain in most tasks.